Hierarchical Pitman-Yor and Dirichlet Process for Language Model

Jen-Tzung Chien and Ying-Lan Chang

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 30010, ROC
jtchien@nctu.edu.tw

Abstract

This paper presents a nonparametric interpretation for modern language model based on the hierarchical Pitman-Yor and Dirichlet (HPYD) process. We propose the HPYD language model (HPYD-LM) which flexibly conducts backoff smoothing and topic clustering through Bayesian nonparametric learning. The nonparametric priors of backoff n-grams and latent topics are tightly coupled in a compound process. A hybrid probability measure is drawn to build the smoothed topic-based LM. The model structure is automatically determined from training data. A new Chinese restaurant scenario is proposed to implement HPYD-LM via Gibbs sampling. This process reflects the power-law property and extracts the semantic topics from natural language. The superiority of HPYD-LM to the related LMs is demonstrated by the experiments on different corpora in terms of perplexity and word error rate.

Index Terms: language model, backoff model, topic model, Bayesian learning

1. Introduction

Statistical language model (LM) plays an important role in many information systems including machine translation, document classification, writing correction, bio-informatics, and speech recognition. The LM \( p(W) \) based on \( n \)-gram aims to calculate the probability of a word string \( W \) by multiplying the probabilities of a predicted word \( w \) conditional on its preceding \( n - 1 \) words. In general, \( n \)-gram model suffers from the inadequacies of training data and long-distance information [7][16]. The modified Kneser-Ney (MKN) LM [6][12] was proposed to tackle the inadequate training data by recursively performing backoff scheme and interpolating with \((n-1)\)-grams. The backoff could be also conducted through a structural Bayesian modeling [25]. To compensate insufficient long-distance information, the topic-based language model [8][9][20][23] was constructed by combining large-span latent topic information [1]. An unsupervised LM adaptation was proposed to incorporate topic mixtures based on latent Dirichlet allocation (LDA) [2].

More recently, Bayesian nonparametric (BNP) learning [3] has been extensively studied in machine learning community. BNP methods flexibly infer the model complexity from data without assuming parametric prior and posterior distributions. Teh [22] proposed a BNP approach to backoff LM according to a hierarchical Pitman-Yor (PY) process [15]. Hierarchical PY (HPY) process draws the power-law distributions which is a striking property of natural languages [10]. HPY-LM was interpreted as the Bayesian extension of MKN-LM [11][22]. In [19], the class-based HPY-LM was established to characterize many-to-many mapping between words and classes for conversational speech. In [24], a doubly HPY-LM was proposed for LM adaptation. A shared LM was adapted to each domain which was represented by an individual HPY-LM. In [14], a nested HPY-LM was combined with dynamic programming for word segmentation. However, these HPY-LMs [11][14][22][24] did not explore topic information. In [17], a PY topic model was constructed but only for document modeling. In [13], HPY process was combined with a topic model for phrase modeling where the parametric LDA model was considered.

This paper presents the BNP learning for LM which allows model growing structurally as more data are observed. We propose a topic-based LM [5] according to the HPY process compound hierarchical Dirichlet process (HDP) [21]. Using this HPYD-LM, the integrated nonparametric priors are constructed to draw topic-dependent backoff \( n \)-grams and simultaneously combine them into a mixture model of topical \( n \)-grams. HDP and HPY are tightly integrated to draw HPYD-LM with power-law property and coherent topic information.

2. Prior Works

2.1. Topic-based language model

Topic-based LM [9] was proposed to capture long-range word dependencies through discovery of latent topics. The resulting \( n \)-gram is expressed by

\[
p(w_i|w_{i-n+1}^{i-1}) = \sum_z p(z_i|w_{i-n+1}^{i-1}) p(w_i|z_i) \tag{1}\]

where \( z_i = k \) denotes the topic label of word \( w_i \) from \( K \) topics and \( p(z_i|w_{i-n+1}^{i-1}) \) denotes the topic proportion given history words \( k = w_{i-n+1}^{i-n} = \{w_{i-n+1}, \ldots, w_{i-1}\} \). The topic-based unigrams and bigrams are calculated by

\[p(w_i) = \sum_z p(w_i|z_i)p(z_i) \quad \text{and} \quad p(w_i|w_{i-1}) = \sum_z p(w_i|w_{i-1}, z_i)p(z_i|w_{i-1}) \], respectively. The maximum likelihood estimates of topic-based LM are calculated according to the expectation-maximization (EM) algorithm. In [8], a cache Dirichlet class LM (cDC-LM) was estimated by a variational Bayes EM procedure where the lower bound of log marginal likelihood was maximized. The likelihood was marginalized over latent classes or topics which were represented by Dirichlet distributions. Different from class-based LM [4] based on hard-clustering, the topic-based LMs [8][9] perform soft-clustering over all topics. Nevertheless, these methods calculated the \textit{parametric mixture models} where the number of topics \( K \) was fixed. BNP learning aims to relax this assumption and conduct the structural learning.

2.2. Bayesian nonparametric learning

BNP learning using HPY [15][22] and HDP [21] have been proposed to infer LM and document model, respectively. HPY process [22] was developed to draw nonparametric \( n \)-gram
model. Given a context $U$ consisting of a sequence of up to $n-1$ history words, HPY-LM calculates the probability of current word $w$ which is sampled from a recursive HPY process $H_U \sim [H_U(w)]_{w \in \Omega_U}$

$$H_u \sim \text{PY}(\theta_0, d_0, H_0), \quad H_U \sim \text{PY}(\theta_{U|1}, d_{U|1}, H_{u(U|1)}) \quad (2)$$

where $H_0$ denotes the word probability over current word $w$ given the empty context $\emptyset$. The global base measure $H_0$ is viewed as a mean vector given by a uniform value $H_0(w) = 1/|\Omega_u|$ for all vocabulary words $w \in \Omega_u$. The prior process $H_u = [H_u(w)]_{w \in \Omega_u}$ draws the unigram $p(w)$. Here, $\theta_0$ and $d_0$ denote the strength parameter and discount parameter, respectively. In case of $d_0 = 0$, $\text{PY}(\theta_0, d_0, H_0)$ is reduced to $\text{DP}(\theta_0, H_0)$ [22]. Moreover, the probability measure $H_U$ over $U$ is drawn from a PY process based on a prior $H_{u(U|1)}$ from backoff context $\pi(U)$. The strength parameter $\theta_{U|1}$ and discount parameter $d_{U|1}$ depend on the length of context $|U|$. Similarly, the backoff measure $H_{u(U|1)}$ is drawn by the same PY process based on its base measure $H_u(\pi(U|1))$ from a even smaller backoff context $\pi(U)$. Given the unigram probabilities $H_0$, a recursive backoff process is implemented to draw probability measures $H_U$ for bigrams, trigrams, etc. In HPY-LM, many unique words are sampled and most of them rarely. Some frequent but rare words are sampled to meet the rich-gets-richer property [22]. Such power-law property is held due to the discount parameters $d_0$ and $d_{U|1}$.

On the other hand, HDP deals with the representation of grouped data where each group is associated with a mixture model. Data in different groups share a global mixture model. The document distribution $H_d$ is drawn from a Dirichlet process (DP) $G_d$, which determines how much a mixture component from a shared mixture model contributes to that document. The dependent unigram $p(w)$ acting as a prior basis for a DP to draw document $H_d$ is different from Bayesian class-based LM [3]. The number of topics $z_k$ is drawn from a recursive HPY process $H_{\pi(U|z)} \sim \text{PY}(\pi(U|z), d_{U|1}, H_{\pi(U|z)})$. The document distribution $H_d$ is reduced to $\text{DP}(\gamma_0, G_0)$ and $\text{DP}(\theta, G_0)$, respectively and $\text{DP}(\gamma_0, G_0)$.

$G_d \sim \text{DP}(\gamma_0, G_0), \quad G_d \sim \text{DP}(\theta, G_0) \quad (3)$

where $\gamma_0$ and $G_0$ denote the strength parameter and base measure of $G_d$, respectively. HDP is developed to represent "a bag of words" from a set of documents through nonparametric prior $G_d$. The sequence of words is not characterized by HDP.

## 3. HPYD Language Model

This paper presents a hierarchical Pitman-Yor and Dirichlet language model (HPYD-LM) which jointly conducts backoff smoothing and topic modeling through BNP learning. The number of topics $K$ is learnt from data. This model is different from Bayesian class-based LM $p(w|h) = \sum_c p(w|h,c)p(c|h)$ [19] where two HPY processes were developed to separately sample mixture probabilities $p(w|h,c)$ and $p(c|h)$ while number of classes $c$ was fixed. In what follows, HPDY process is shown as a single compound process.

### 3.1. HPYD process

HPYD-LM assumes that $n$-gram is expressed by a nonparametric topic mixture model. HPYD process is described as follows. Starting from the uniform seed measure $H_0$, we draw a global topic mixture measure $G_0 \sim \text{DP}(\gamma_0, G_0)$, which is sampled by $H_{U|1} \sim \text{PY}(\theta_1, d_1, G_0)$ where $G_0$ is acted as a prior base measure. Next, $H_{U|z_i}$ serves as a base measure for a DP to draw unigram probability $G_{w|1} \sim \text{DP}(\gamma_1, H_{U|z_i})$. Using $G_{w|1}$ as a prior measure, we draw topic-dependent bigram by using PY process $H_{U|w_{i-1}z_i} \sim \text{PY}(\theta_2, d_2, G_{w|1})$. This measure is again acted as a prior basis for a DP to draw bigram $G_{w_{i-1}w_{i} \mid 1} \sim \text{DP}(\gamma_2, H_{U|w_{i-1}z_i})$. Using bigram measure $G_{w_{i-1}w_{i} \mid 1}$ as a basis, the topic-dependent trigram is drawn by a PY process $H_{U|w_{i-1}w_{i}z_i} \sim \text{PY}(\theta_3, d_3, G_{w_{i-1}w_{i} \mid 1})$. Having the prior measure $H_{U|w_{i-1}w_{i}z_i} \sim \text{DP}(\gamma_3, H_{U|w_{i-1}w_{i}z_i})$, topic-dependent trigram probability is drawn by $G_{w_{i-1}w_{i}w_{i+1} \mid 1} \sim \text{DP}(\gamma_3, H_{U|w_{i-1}w_{i}z_i})$. Therefore, HPYD process is recursively realized by sampling topic-dependent $n$-gram probability $p(w_{i|w_{i-1}z_i})$ and then $n$-gram probability $p(w_{i|w_{i-1}z_i})$.

$$H_{U|w_{i-1}w_{i}z_i} \sim \text{PY}(\theta_3, d_3, G_{w_{i-1}w_{i} \mid 1}) \quad G_{w_{i-1}w_{i}w_{i+1} \mid 1} \sim \text{DP}(\gamma_3, H_{U|w_{i-1}w_{i}z_i}) \quad (4)$$

where $\theta_1, \ldots, \theta_n$ and $d_1, \ldots, d_n$ denote the strength and discount parameters of PY process, respectively and $\gamma_0, \ldots, \gamma_n$ denote the strength parameters of DP.

### 3.2. New Chinese restaurant scenario

We implement the nonparametric solution to HPYD-LM in (4) through a new Chinese restaurant process as illustrated in Figure 1. Imagine that there are Chinese restaurants serving customers with infinite tables (yellow) $l$, infinite menus (blue) $k$ and infinite dishes (green) $i$. For each restaurant (red) or context $U$, the first customer or word $x_1$ enters the restaurant, sits with the first table, and draws a single menu shared for the customers in the same table. He/she orders a dish by this menu. As shown by arrows, tables 1 and 3 draw the same menu table 2 draws menu 2 and table 4 draws menu 3. Let $c_{ul}$ denote the number of customers in table $t$ and $n_{uk} \in \mathbb{N}$ denote the number of customers ordering dish $l$ which is labelled by a distinct word $w$ from menu $k$ given context $U$. We have $c_{ul} = \sum_l c_{ul}$, $n_{uk} = \sum_l n_{uk}$, $n_{uk} = \sum_l c_{ul}$. Let $m_{uk}$ denote the number of dishes in menu $k$ which are labelled by distinct word $w$. Number of occupied tables in restaurant $U$ is expressed by $m_{u}$. According to this metaphor, each table $l$ is associated with a topic for a distinct menu $k$ and each dish $i$ is associated with an $n$-gram for a distinct word $w$. More details are given below.

The $i$th customer $x_i$ enters a restaurant $U$ and selects either an occupied table with probability $\frac{c_{ul}}{c_{ul} + \gamma_i}$ or a new table

![Figure 1: Chinese restaurant scenario for HPYD-LM.](image-url)
with probability $\frac{\gamma|U|}{c_u + \gamma|U|}$. For a new table, this customer either
draws an existing menu $k$ with the probability $\frac{m_k}{n_u + \gamma|U|}$ or a new
menu with probability $\frac{\gamma|U|}{n_u + \gamma|U|}$. The number of tables drawing
menu $k$ in all restaurants $m_k$ is used. Different tables may
choose the same menu. After selecting a table with menu $k$, the
customer $x_t$ further selects either an ordered dish $l$ with proba-
bility $\frac{\gamma x_t u_k \pi_{U|l}}{n_u + \gamma|U|}$, or a new dish with proba-
bility $\frac{\gamma x_t u_k \pi_{U|l}}{n_u + \gamma|U|} p(\pi_{U|l})$ for each dish.
The dishes for $(n-1)$-gram come from back-off restaur-
ten $\pi(U)$ with measure $p(\pi_{U|l})$. Number
of dishes $\lambda_{w|w}$ is counted over different menus. Combining
with new dish provides the approach to model smoothing.

### 3.3. Gibb’s sampling inference for HPYD-LM

HPYD-LM is inferred according to Gibb’s sampling based on
this new Chinese restaurant franchise. First, the general topic
measure is implemented as a mixture model for $K$ menus (or
topics), i.e. $G_\gamma \sim \sum_{k=1}^{N} \frac{n_{uk}}{n_u + \gamma} \delta_{\theta_k} + \frac{\gamma}{n_u + \gamma} H_0$ where $H_0$
denotes the atom of topic mixture model for menu $k$. Next,
we draw topic-dependent unigram $H_0(z=k)$ for a word $w$
by considering $G_\gamma$ as a base measure according to the
PY process. With the prior measure $H_0(z=k)$, we draw unigram $G_{n|w}$
by a DP. In this fashion, HPYD $n$-gram with context $U = w_{i-n+1}$
is recursively sampled by

$$
H_{n|w_{i-1}} = \frac{n_{ukw}}{n_{uk} + \theta} + \frac{\theta n_{uw} - \theta \lambda_{uw}}{\theta + \theta n_{uw}} G_{n|w_{i-1}}
$$

$$
G_{n|w_{i-1}} \sim \sum_{k=1}^{K} n_{ukw} \cdot H_{n|w_{i-1}}(z=k)
$$

which is realized from (4). In (5), the topic-dependent $n$-gram
$H_{n|w_{i-1}}(z=k)$ is first drawn by a PY process with a prior
$G_{n|w_{i-1}}(z=k)$ from backoff context $\pi(U) = w_{i-n+1}$, and subse-
sequently treated as a prior to draw $n$-gram $G_{n|w_{i-1}}(z=k)$
through a DP. Using this HPYD-LM, backoff weights depend on the
number of dishes $\lambda_{ukw}$ labelled by word $w_i$ in a menu $k$. The
topic-dependent $n$-grams are determined through drawing the
dishes from different contexts. Topic proportion is decided by
the number of customers $c_{n|w_{i-1}}(z=k)$ sitting in the tables which or-
der the same menu $k$. Latent topics are autonomously produced
by choosing new menus. Therefore, considering the topic-based
LM in (1) and the sitting and ordering arrangements of tables,
menus and dishes, we infer the nonparametric HPYD-LM
$p(w_t|w_{i-1}, z_i)$ which is proportional to

$$
\prod_{k=1}^{K} \sum_{c_{n+w_i}} c_{n+w_i} \left[ \frac{n_{ukw}}{n_{uk} + \theta} + \frac{\theta n_{uw} - \theta \lambda_{uw}}{\theta + \theta n_{uw}} \right] \cdot p(w_t|w_{i-1}, z_i = k)
$$

(6)

Notably, (6) is viewed as a mixture model consisting of $K$ ex-
tisting topic-dependent $n$-grams shown in brackets and a possibly-
generated new mixture given in the second term.

We apply the Gibb’s sampling algorithm to sample tables,
menus and dishes from training data according to the conditional
posterior distributions $p(l|t_i, z_i, w, U)$, $p(z_i = k|t_i, z_i, w, U)$, respectively,
where $w = \{w_1, w_{i-1}\}$, $t = \{t_i, t_{i-1}\}$, $z = \{z_i, z_{i-1}\}$, and "$-$
" denotes the self-exception. The setting arrangement is deter-
mimed by sampling table $t$ according to either $p(l|t_i, z_i, w, U)$
which is proportional to $c_{n+i} \cdot p(w_t|t_i, z_i, w, U, l)$ if table $t$
is occupied or $\gamma n \cdot p(w_t|l_i = new, t_i, z_i, w, U)$ if table is new.
After setting in a new table, we sample a distinct menu or topic
for this table given context $U$ by $p(x_t = k|x_i = k, z_i, t_i, w, U)$
which is proportional to either $m_k \cdot p(x_t = k, z_i, w, U)$ if
menu $k$ is ordered or $\gamma \cdot p(x_t = k, z_i, w, U)$ if menu $k$ is new. Next, we draw a dish in menu $k$ by $p(l_i|z_i = k, z_i, w, U)$
which is proportional to either $\max\{n_{ukw}, - d_{U|l_i}\} \pi_{U|l_i}$
if dish $l_i$ is ordered or $\theta_{U|l_i} + d_{U|l_i} \Delta G_{n|w_i|l_i}(U)$
if dish $l_i$ is new. The counts $c_{n+i}$ and $n_{ukw}$ are measured
over all words except $w_i$.

### 4. Experiments

#### 4.1. Experimental setup

We evaluate the proposed HPYD-LM by using three datasets
with different contents and data sizes. The metrics of perplex-
ity and word error rate (WER) (%) are evaluated. In continu-
ous speech recognition, we adopted the Wall Street Journal
(WSJ) 1987-1989 corpus containing 86K documents with 38M
words and a vocabulary size of 5000. A total of 330 test sen-
tences were sampled from November 1992 ARPA CSR benchmark
data. The SI-84 training set was used to estimate HMM parameters
based on 39-dimensional MFCC feature vectors.

System configuration was detailed in [8]. Two other datasets
were collected for evaluation of perplexity. First, the Associa-
ted Press newswire (AP) 1989 dataset consisted of 84,778
documents and 1,706,742 sentences with a vocabulary size of
16003. AP was partitioned into a training set with 36,727,391
words and a test set with 4,022,423 words. Second, NIPS09.12
(http://arbanly.net/resources.html) contained 1740 papers from
NIPS conferences. We collected a total of 2,034,215 words with
a vocabulary size of 3360. NIPS papers were divided into a
training set with 1,830,392 words and a test set with 203,823
words.

For comparative study, we carried out trigram LM’s by using
LDA-LM [20], cDC-LM [8], MKN-LM [6][12] and HPY-LM
[11][22]. MKN-LM was carried out by using [18]. The results
of LDA-LM and cDC-LM were obtained by interpolating with
MKN-LM. HPY-LM and HPY-DLM were implemented by per-
forming Gibbs sampling with 200 iterations and 100 samples at
each iteration. The burn-in samples in the first 20 iterations
were abandoned. Representative samples from the stationary
distribution were collected. The parameters $d_{U|l_i} \sim \Gamma(2, 5)$
and $\theta_{U|l_i} \sim \Gamma(1, 10)$ were drawn with randomly selected
and $\beta$. The parameter $\gamma_{U|l_i} = 100$ was fixed for all contexts
$U$ and $n$-grams. Perplexity of test data was examined by using
SRI toolkit [18] for comparison. We found that the larger the
number of words choosing a specific topic, the higher the aver-
age discount based on $\lambda_{ukw}$ is calculated. This phenomenon
meets the power law property.

#### 4.2. Experimental results

First of all, Figure 2 displays the estimated topic proportions
(number of customers in the tables) in different topics (menus)
at different sampling iterations. WSJ corpus is used. The aver-
age log probability and the number of estimated topics are
shown. Gibbs sampling converges in these iterations. At each
iteration, the topics are adaptively estimated. Only a small num-
ber of topics have large topic proportions while many topics
have small topic proportions. Distribution of topic proportions conforms with power-law property. Table 1 shows an example of topic words from five selected topics which are extracted via HPYD-LM using WSJ corpus. It is obvious that topic words within a latent topic are semantically similar while those across topics are significantly different. This topic information is beneficial for estimation of language model. Using HPYD-LM, LMs under different size of training data.

Figure 3 displays the perplexity versus the number of topics or classes by using LDA-LM, cDC-LM, HPY-LM and the proposed HPYD-LM. WSJ corpus is adopted. Using LDA-LM and cDC-LM, the parametric topic priors are introduced in Bayesian topic-based LMs where the number of topics is fixed to be $K=20, 50, 100, 150$ and $200$. We can see that perplexity is reduced by increasing number of topics. The lowest perplexity of LDA-LM $(114.10)$ is achieved by using 150 topics. However, BNP learning based on HPY-LM and HPYD-LM reduces the perplexity as 112.20 and 106.84, respectively. Using HPYD-LM, the number of topics (60) is automatically determined. Nonparametric topic priors are introduced to build an effective and compact topic-based LM. Furthermore, Table 2 lists the perplexities of MKN-LM, LDA-LM, cDC-LM, HPY-LM and HPYD-LM by using AP and NIPS datasets. HPYD-LM achieves the lowest perplexity among these methods. In case of AP dataset, MKN-LM, LDA-LM, cDC-LM, HPY-LM and HPYD-LM obtain the perplexity of 115.82, 112.35, 109.11, 110.07 and 97.81, respectively. On the other hand, we investigate the performance of continuous speech recognition by using different LMs as shown in Table 3. In this comparison, WERs are reported under comparable model complexity. The results of LDA-LM and cDC-LM were implemented by fixing number of topics or classes to be 60. In this table, HPYD-LM and HPYD-LM2 imply the HPYD-LM with hyperparameters $\gamma_0 = 1$ and $\gamma_0 = 10$ and lead to 52 and 60 topics in the estimated HPYD-LMs, respectively. In this set of experiments, LDA-LM and cDC-LM are interpolated with MKN-LM and outperform MKN-LM. CDC-LM, HPY-LM obtain the perplexity of 115.82, 112.35, 109.11, 110.07 and 97.81, respectively. On the other hand, we investigate the performance of continuous speech recognition by using different LMs as shown in Table 3. In this comparison, WERs are reported under comparable model complexity. The results of LDA-LM and cDC-LM were implemented by fixing number of topics or classes to be 60. In this table, HPYD-LM and HPYD-LM2 imply the HPYD-LM with hyperparameters $\gamma_0 = 1$ and $\gamma_0 = 10$ and lead to 52 and 60 topics in the estimated HPYD-LMs, respectively. In this set of experiments, LDA-LM and cDC-LM are interpolated with MKN-LM and outperform MKN-LM. CDC-LM, HPY-LM obtain the perplexity of 115.82, 112.35, 109.11, 110.07 and 97.81, respectively. Among these LMs, the lowest WER 4.82% is achieved by using HPYD-LM with $\gamma_0 = 10$. HPYD-LM could estimate compact LM for speech recognition. The improvement using HPYD-LM comes from the flexibility of backoff smoothing and the contribution of scalable topic information.

Table 1: Topic words using HPYD-LM under different topics.

<table>
<thead>
<tr>
<th>Trade</th>
<th>Investment</th>
<th>Stock Market</th>
<th>Politics</th>
<th>Economics</th>
</tr>
</thead>
<tbody>
<tr>
<td>finance</td>
<td>offer</td>
<td>plan</td>
<td>public</td>
<td>initial</td>
</tr>
<tr>
<td>company</td>
<td>previously</td>
<td>agency</td>
<td>secretary</td>
<td>information</td>
</tr>
<tr>
<td>threat</td>
<td>talk</td>
<td>administration</td>
<td>strike</td>
<td>yen</td>
</tr>
<tr>
<td>meeting</td>
<td>interview</td>
<td>bank</td>
<td>Bush</td>
<td>international</td>
</tr>
<tr>
<td>transaction</td>
<td>party</td>
<td>dollar</td>
<td>accepted</td>
<td>institute</td>
</tr>
<tr>
<td>timing</td>
<td>debt</td>
<td>cash</td>
<td>population</td>
<td>insurance</td>
</tr>
<tr>
<td>thrift</td>
<td>investment</td>
<td>Dow</td>
<td>aggregate</td>
<td>investment</td>
</tr>
<tr>
<td>billion</td>
<td>treasury</td>
<td>stock</td>
<td>tax</td>
<td>inflation</td>
</tr>
<tr>
<td>trade</td>
<td>bank</td>
<td>stock</td>
<td>favorable</td>
<td>income</td>
</tr>
</tbody>
</table>

Table 3: WER (%) versus different LMs and hyperparameters.

<table>
<thead>
<tr>
<th></th>
<th>MKN-LM</th>
<th>LDA-LM</th>
<th>cDC-LM</th>
<th>HPY-LM</th>
<th>HPYD-LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>115.82</td>
<td>112.35</td>
<td>109.11</td>
<td>110.07</td>
<td>97.81</td>
</tr>
<tr>
<td>NIPS</td>
<td>163.18</td>
<td>159.75</td>
<td>155.27</td>
<td>153.02</td>
<td>148.05</td>
</tr>
</tbody>
</table>

5. Conclusions

We presented the HPYD-LM based on a new random process which combined HPY for constructing the topic-dependent backoff smoothed LMs and HDP for integrating these LMs into a topic mixture model. Model selection issue was tackled by flexibly extending the number of topics. A Chinese restaurant franchise was proposed to implement the HPYD-LM which satisfied the properties for power-law distribution and topic mixture distribution. The posterior probabilities for drawing tables, menus and dishes were derived. Gibbs sampling was applied to infer HPYD-LM parameters. The experiments on WSJ, AP and NIPS showed that HPYD-LM outperformed the other LMs in terms of perplexity and word error rate.
6. References


